INTRODUCCIÓ A XML

UF 1: PROGRAMACIÓ AMB XML - PART 1

CARLES GONZÀLEZ

DADES

- Els humans utilitzen les dades per viure
- Les dades son una representació d'aspectes del món real
- Generalment les dades necessiten algun tipus de procés per poder ser usades:
  - Es poden fer servir per fer càlculs o per prendre decisions
  - El procés les converteix en 'informació'

REPRESENTACIÓ DE LES DADES

- Els ordinadors emmagatzemen aquestes dades en binari a través d'algun tipus de codificació:
  - Text: ASCII, UTF-8, ISO-8859-15, Windows- 1251, etc...
  - Binària:
    - Imatges (JPG, PNG, GIF, ...)
    - So (WAV, MP3, ...)
    - Vídeo (MPEG, ...)
    - Etc..

ESTRUCTURACIÓ DE DADES

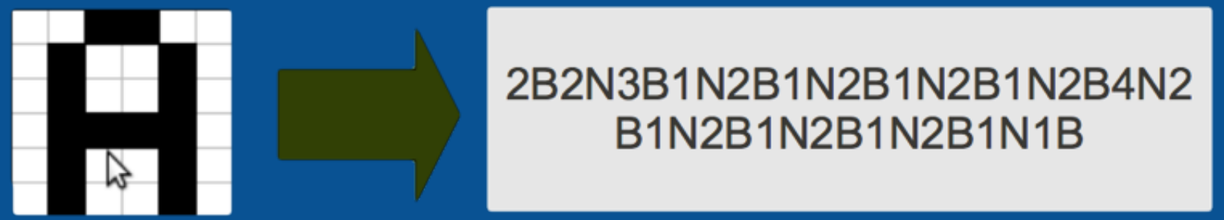
- Les dades en els ordinadors es guarden en fitxers
- Tradicionalment els fitxers s'emmagatzemen en en dos tipus de formats:
  - Fitxers binaris
  - Fitxers de text

FITXERS BINARIS

- Els fitxers binaris són simplement una tira de bits
- La informació que contenen només la entenen els programes que l'han generat
  - Per això els fitxers binaris només es poden llegir amb els programes que els han creat
- Els fitxers binaris estan molt bé perquè són llegits fàcilment pels ordinadors
  - Les dades es llegeixen i guarden d'una forma molt eficient

FITXERS BINARIS

» A vegades els fitxers binaris contenen informació sobre el seu contingut per diferents motius



» Algú ho entendria sense informació?

» Si el creador no ho explica és complicat

» En l'exemple els números són "metadades"

METADADES

“Les metadades són dades sobre les dades”

FITXERS DE TEXT

- Els fitxers de text també són tires de bits però estan agrupats de forma estandarditzada
  - ASCII, UTF-8, UTF-16, EBCDIC, Windows-1250
- Gràcies a l'estàndard aquests fitxers es poden obrir en diferents programes
- La informació es comparteix més fàcilment amb fitxers de text que amb fitxers binaris

FITXERS DE TEXT

Desavantatges:

- Els sistemes operatius tracten de forma diferent alguns dels seus aspectes (com els salts de línia)
- És complicat afegir-hi informació sobre les dades, metadades, ja que no seran interpretades
  - Poden arruïnar-ne totalment la facilitat de lectura
  - No existeix una forma estàndard d'afegir-les

FITXERS DE TEXT

» Antigament per representar dades es feia separant els valors amb comes o algun altre símbol (csv, tsv, ...)

```
"Nom","Cognom","Ofici","Naixement","Poblacio","Punts"
"Filomenu","Garcia","Professor", "10/04/1902","Cabanes",12
"Mariano","Puigdevall","Informàtic","19/05/1972","Cabanes",23
"Federicu", "Pi", "Mestre","20/03/1968","Girona",40
```

» S'ha de saber que la primera línia són metadades

» Afegir-hi noves dades pot ser molt problemàtic pel programa que les llegeixi

» Probablement haurem de canviar el programa

FITXERS DE MARQUES

- Els llenguatges de marques recullen el millor dels dos tipus de fitxers:
  - Dels fitxers binaris
    - La facilitat de posar metadades en el contingut
  - Dels fitxers de text:
    - La facilitat d'intercanvi d'informació
    - L'estandardització

FITXERS DE MARQUES

- Els llenguatges de marques estan basats en text
  - Poden ser creats amb qualsevol editor de textos
  - Però no estan pensats per ser llegits

LLENGUATGES DE MARQUES

» Un llenguatge de marques combina dades i etiquetes que les marquen i que contenen informació addicional sobre l'estructura del text o la seva presentació.

» Les marques estan barrejades amb el propi text.

```
<persona>
    <nom>Pepe</nom>
    <cognom>Pérez</cognom>
</persona>
```

LLENGUATGES DE MARQUES

- Tot i que els sistemes de marques en que ens concentrarem són els d'estil "web" cal no oblidar que n'hi ha d'altres:
  - Documentació
    - Wikitext, TeX, DocBook
  - Text enriquit
    - RTF
  - Intercanvi de dades
    - XML, JSON, LDIF

LLENGUATGES DE MARQUES

JSON

```
{
 "persona": {
   "nom": "Pepe",
   "cognom": "Pérez",
   "telefon": "+1 408 555 8585"
 }
}
```

LDIF

```
givenname: Pepe
cn: Pérez
telephonenumber: +1 408 555 8585
```

LLENGUATGES DE MARQUES

» El llenguatge de marques més conegut és l'HTML

» És el que es fa servir en les pàgines web

» Però no és el primer que ha existit, ni l'únic

```
<html>
  <head>
    <title>
      Pàgina
    </title>
  </head>
  <body>
    Hola!
  </body>
</html>
```

SGML

» La primera tecnologia estandarditzada de llenguatges de marques va ser l'SGML

» Es va fer servir com estàndard de la informació de propòsit general

» Partia de la idea de que s'han de separar les dades d'un document de la seva forma

PROBLEMES SGML

» La majoria dels documents estaven destinats a la impressió

» Terriblement complex de manera que només el feien servir els especialistes

HTML

» El 1989, Tim Berners-Lee i Anders Berglund, dos investigadors del CERT, van crear un llenguatge basat en etiquetes destinat a compartir informació per Internet: HTML

» HTML és un format que descriu la visualització d'una pàgina web

» HTML està molt orientat a la visualització

TECNOLOGIA WEB

- HTML ha tingut un èxit extraordinari i molt ràpid
- Això ha fet que les tecnologies web no parin d'evolucionar
- HTML ha sofert molts canvis al llarg dels anys
- El suport HTML dels navegadors cada vegada és més complexe
  - Les pàgines HTML no sempre es veuen igual en els diferents navegadors

TECNOLOGIES WEB

- L'HTML és molt difícil de reutilitzar
  - És molt difícil representar-hi informació que es pugi reutilitzar en altres llocs
    - Poder presentar la informació de diferents formes
    - Personalitzar les dades
  - Fa falta alguna forma de poder fer-hi recerques intel·ligents i seleccionar-ne el resultats

TECNOLOGIES WEB

```
<html>
  <head>
    <title>
      Professor
    </title>
  </head>
  <body>
    <p>
      Nom: Federicu Pi
    </p>
  </body>
</html>
```

“Com pot una màquina determinar automàticament què és el nom, què el cognom, ...?”

TECNOLOGIES WEB

Feia falta una forma de:

“"Buscar, moure, visualitzar i manipular la informació continguda en els documents HTML"”

NAIXEMENT D'XML

- El consorci W3C va desenvolupar una alternativa a l'HTML que podés satisfer les necessitats futures del web.
- El 1996 el consorci W3C es va proposar introduir el poder i la flexibilitat de l'SGML al web.
- SGML oferia tres avantatges que l'HTML no tenia:
  - Extensibilitat
  - Estructura
  - Validació

ESPECIFICACIONS XML

EXtensible Markup Language

» El febrer de 1998 es llença l'especificació 1.0 d'XML:

  » http://www.w3.org/TR/2004/REC-xml-20040204/

» L'ultima especificació d'XML és la 1.1 que va sortir el 2004:

  » http://www.w3.org/TR/xml11/

Totes les especificacions es revisen periòdicament

QUÈ ÉS XML?

» XML és un simple llenguatge de descripció d'informació

» És un estàndar que permet dissenyar i desenvolupar llenguatges de marques.

» XML és un format de text estandarditzat que serveix per representar i transportar informació estructurada.

PRESENTACIÓ

Una de les idees més importants és:

“Separar les dades de la presentació”

» XML no es preocupa de com es presentaran les dades als usuaris

» Per fer la presentació ja s'han desenvolupat mecanismes:

  » CSS

  » XSL-FO

  » etc...

ETIQUETES

» A HTML li ha anat bé amb un número finit d'etiquetes per la presentació

» D'altra banda, tots els intents per crear un conjunt finit i general d'etiquetes per a descriure les dades van fallar

  » Cada conjunt d'usuaris en necessita un de diferent (matemàtics, químics, etc..

» La solució adoptada va ser la més lògica:

"Un número infinit d'etiquetes"

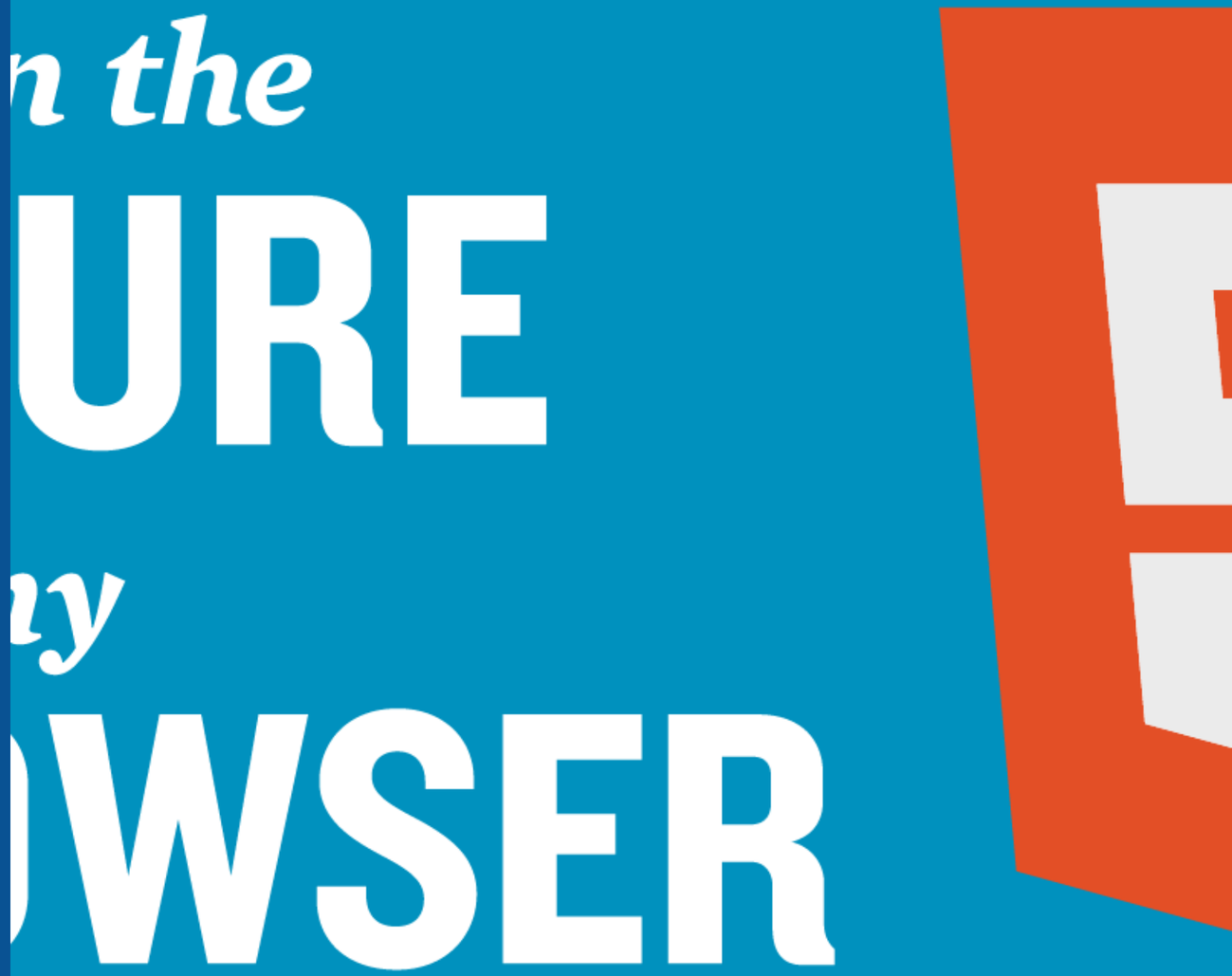
ESTRUCTURACIÓ DE DADES

» Un altre dels objectius és donar una estructura a les dades (Són més fàcils d'interpretar)

```
<!--Els alumnes són: Federicu Garcia, Filomenu Pi,...-->
<modul>
  <nom>
    Llenguatges de marques
  </nom>
  <alumnes>
    <nom>
      Federicu Garcia
    </nom>
    <nom>
      Filomenu Pi
    </nom>
    ...
  </alumnes>
</modul>
```

TRANSPORT DE DADES

- XML i JSON estan pensats per transportar dades
  - A diferència d'HTML4 i XHTML si que es pot determinar de forma automàtica què significuen les dades
- HTML5 permet descriure millor les dades emmagatzemades.



TRANSPORT DE DADES

```
<!--XHTML-->
<html>
  <head>
    <title>Professors</title>
</head>
<body>
  <p>Federicu Pi</p>
  <p>Mariano Po</p>
</body>
</html>
```

```
<!--XML-->
<professors>
  <professor>
    <nom>Federicu</nom>
    <cognom>Po</cognom>
  </professor><professor>
    <nom>Mariano</nom>
    <cognom>Po</cognom>
  </professor>
</professors>
```

Podem respondre:

» Quina informació conté el fitxer?

» Quina és la estructura de la informació?

» Quins tags s'han creat per descriure'n la informació?

FORMATS ESTÀNDARD

- Tenim la capacitat de crear un vocabulari que només entengui el nostre programa
  - No necessita llicència
- O podem fer-lo obert perquè l'entengui tothom
  - Al fer servir el mateix format la comunicació de dades és més fàcil
  - Hi ha molts vocabularis estàndards XML

FORMATS ESTÀNDARD

FORMAT	OBJECTIU
SVG	Pensat per gràfics vectorials escalables 2D
MathML	Representació de fórmules matemàtiques
CML	Intercanvi d'informació química
SMIL	Tractament de la informació multimèdia
SSML	Síntesi de la veu
Comptabilitat	XFRML (Extensible Financial Reporting Markup Language)
Vcard	Intercanvi de contactes
SCORM	Cursos (Moodle, etc...)
ChessGML	Partides escacs

I molts més...

FORMATS ESTÀNDARD

- Molts programes que feien servir formats binaris han passat a algun tipus d'XML:
  - Microsoft Office
    - Va passar de guardar els documents en binari .DOC a XML .DOCX (OOXML) al estandaritzar-lo
  - OpenOffice.org
- Molts dels documents de configuració dels sistemes operatius estan en XML!

```
# Linux
$ locate .xml | wc -l
21829
```

```
# Windows XP
C:\> dir /a-d /s *.xml | find /c /v ""
698
```

EXTENSIBLE

- Un altre dels avantatges de XML és que es fàcilment extensible i adaptable
  - Creem els tags que tinguin significat per nosaltres
  - Podem crear el vocabulari que ens faci falta per allò que busquem
- Hi ha formes de definir quina és la estructura que nosaltres definim
  - Diversos estàndards DTD, XML Schema Language, Relax NG, etc..
  - Ens serviran per comprovar que el document compleix amb les normes del vocabulari

USOS

- XML s'està fent servir en múltiples camps:
  - Contingut de pàgines web
    - Un dels estàndards que es fan servir en pàgines web (XHTML) està basat en XML
    -
    - XML de forma inherent té múltiples formes en que pot ser representat (XSL-FO, CSS, ...)

USOS

- Computació distribuïda
  - L'intercanvi de dades entre sistemes diferents que permetin les crides entre objectes entre màquines
- Comerç electrònic
  - Bussines to Bussines, Bussines to Consumer

PROBLEMES

- XML ocupa més espai a disc que els seus equivalents en format binari
  - Hi ha tendència a crear fitxers molt grans
  - Això pot tenir un impacte en el rendiment dels programes
    - El fitxer és molt gran i en format text
    - Es especialment important quan es transfereix per internet
- Es difícil d'escriure manualment
- Es costós de parsejar

ALTERNATIVES

- JSON
  - Molt més compacte que el XML
  - Fàcil d'escriure i parsejar
  - Apropiat per a la tranferència d'informació entre aplicacions
- YAML
  - JSON augmentat
  - Molt utilitzat en fitxers de configuració